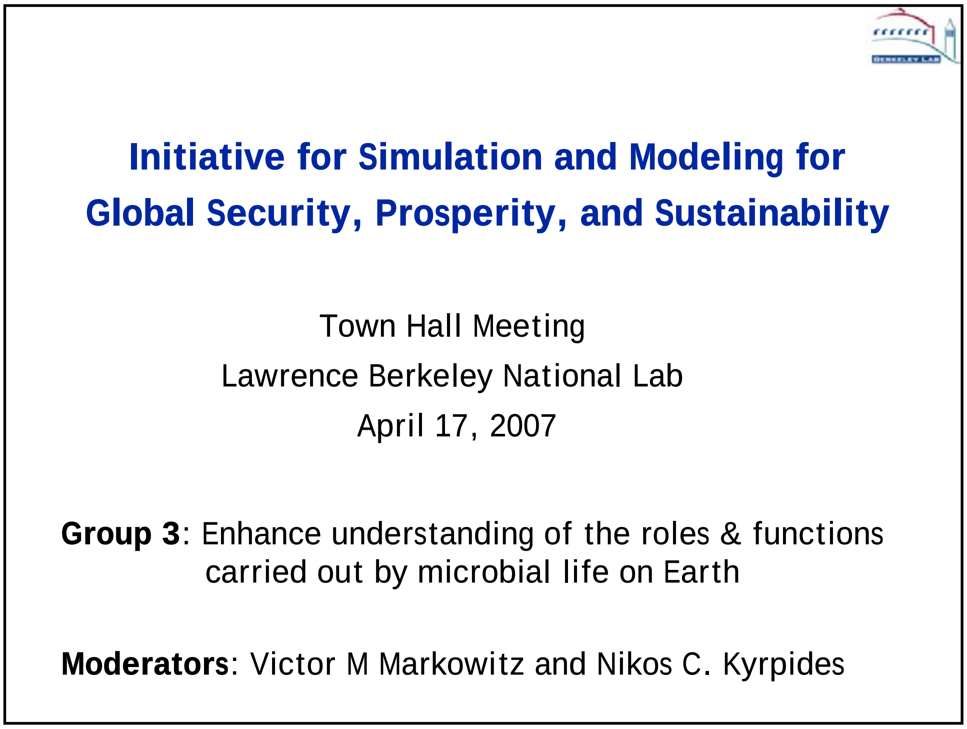
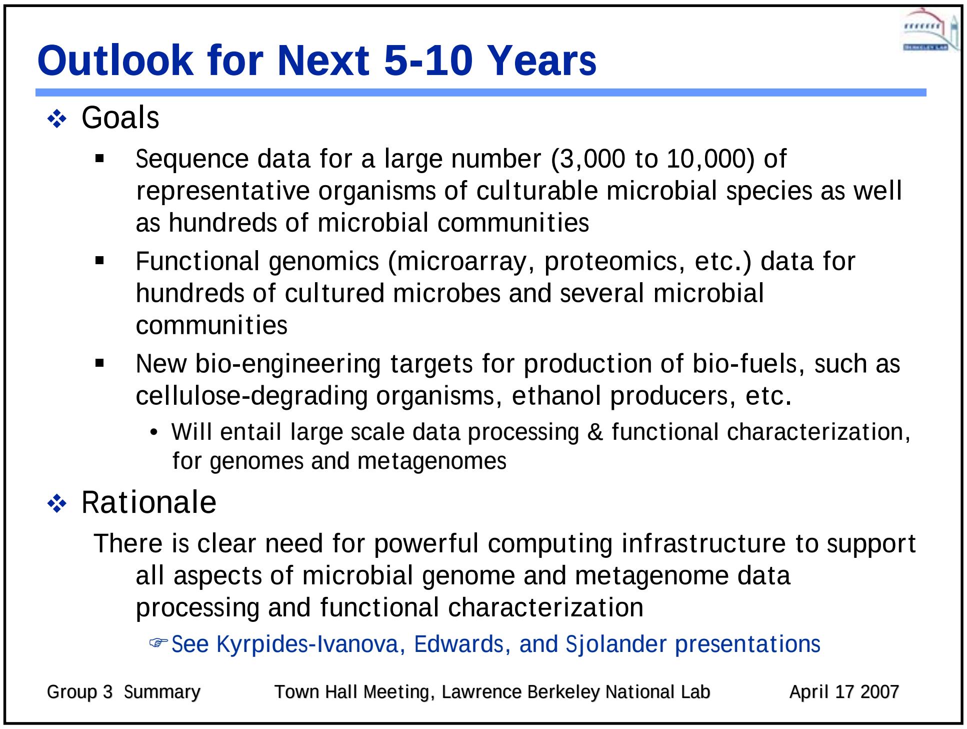


Initiative for Simulation and Modeling for Global Security, Prosperity, and Sustainability

Town Hall Meeting
Lawrence Berkeley National Lab
April 17, 2007

Group 3: Enhance understanding of the roles & functions carried out by microbial life on Earth

Moderators: Victor M Markowitz and Nikos C. Kyrpides

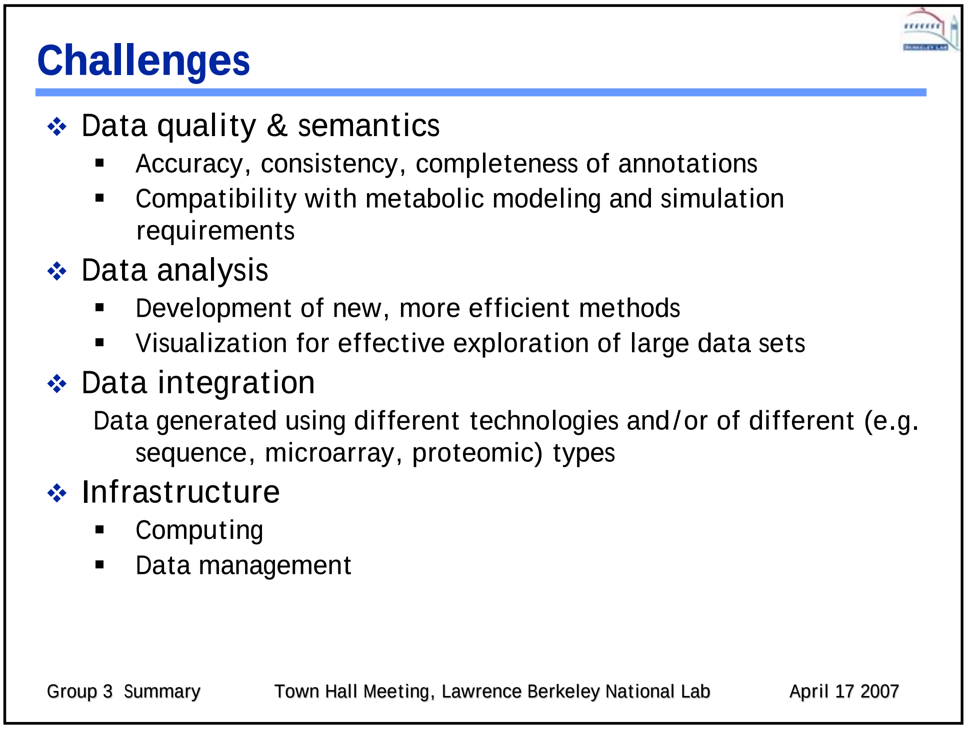


Outlook for Next 5-10 Years

- Goals
  - Sequence data for a large number (3,000 to 10,000) of representative organisms of culturable microbial species as well as hundreds of microbial communities
  - Functional genomics (microarray, proteomics, etc.) data for hundreds of cultured microbes and several microbial communities
  - New bio-engineering targets for production of bio-fuels, such as cellulose-degrading organisms, ethanol producers, etc.
    - Will entail large scale data processing & functional characterization, for genomes and metagenomes
- Rationale

  There is clear need for powerful computing infrastructure to support all aspects of microbial genome and metagenome data processing and functional characterization

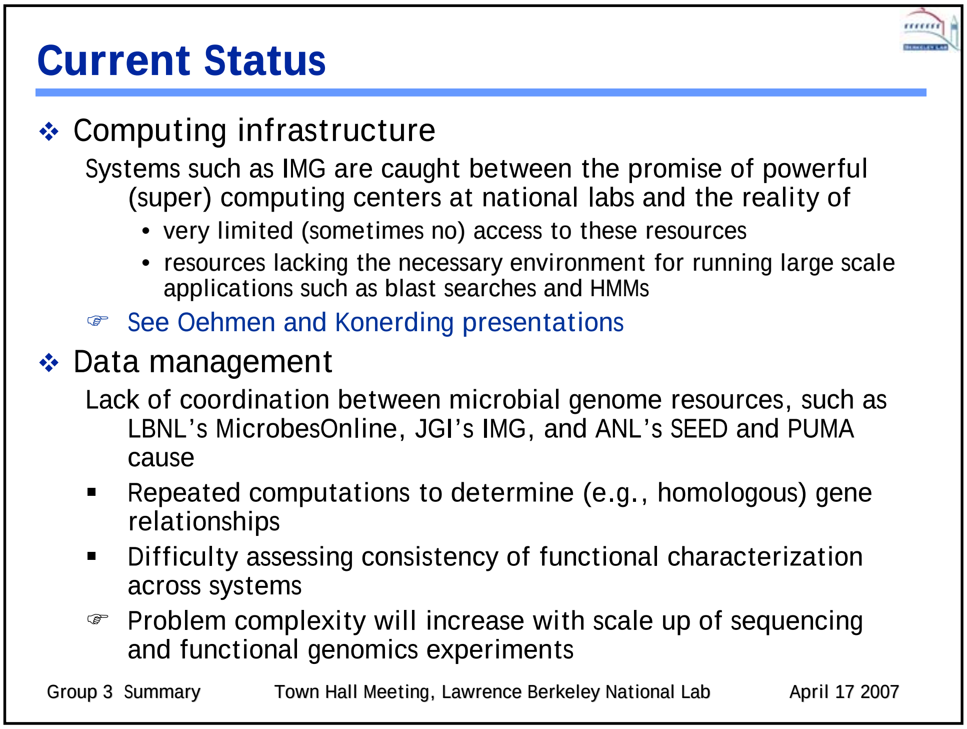
  ☞See Kyrpides-Ivanova, Edwards, and Sjolander presentations



Challenges

- Data quality & semantics
  - Accuracy, consistency, completeness of annotations
  - Compatibility with metabolic modeling and simulation requirements
- Data analysis
  - Development of new, more efficient methods
  - Visualization for effective exploration of large data sets
- Data integration

  Data generated using different technologies and/or of different (e.g. sequence, microarray, proteomic) types
- Infrastructure
  - Computing
  - Data management



Current Status

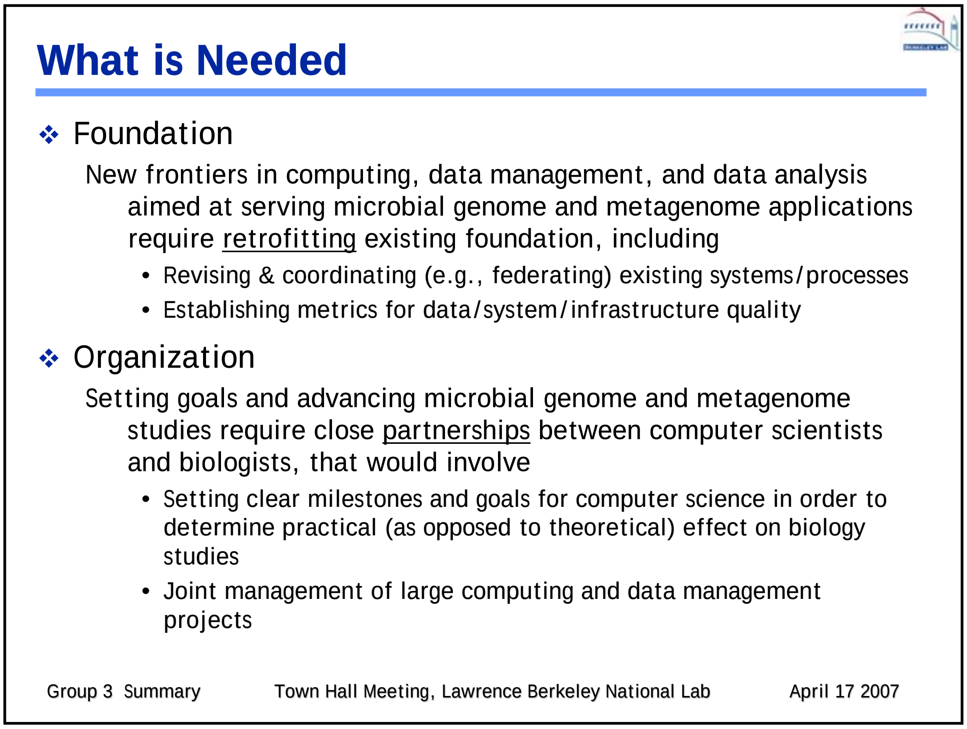
- Computing infrastructure

  Systems such as IMG are caught between the promise of powerful (super) computing centers at national labs and the reality of
  - very limited (sometimes no) access to these resources
  - resources lacking the necessary environment for running large scale applications such as blast searches and HMMs

  ☞ See Oehmen and Konerding presentations
- Data management

  Lack of coordination between microbial genome resources, such as LBNL's MicrobesOnline, JGI's IMG, and ANL's SEED and PUMA cause
  - Repeated computations to determine (e.g., homologous) gene relationships
  - Difficulty assessing consistency of functional characterization across systems

  ☞ Problem complexity will increase with scale up of sequencing and functional genomics experiments



What is Needed

- Foundation

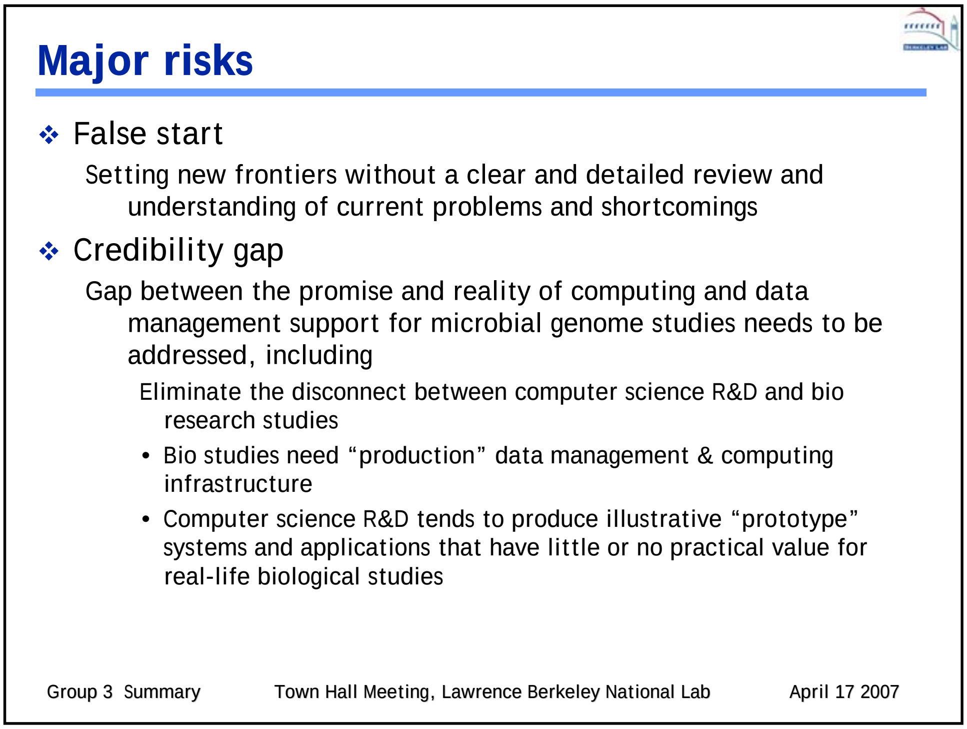
  New frontiers in computing, data management, and data analysis aimed at serving microbial genome and metagenome applications require retrofitting existing foundation, including

  - Revising & coordinating (e.g., federating) existing systems/processes
  - Establishing metrics for data/system/infrastructure quality

- Organization

  Setting goals and advancing microbial genome and metagenome studies require close partnerships between computer scientists and biologists, that would involve

  - Setting clear milestones and goals for computer science in order to determine practical (as opposed to theoretical) effect on biology studies
  - Joint management of large computing and data management projects



Major risks

- False start

  Setting new frontiers without a clear and detailed review and understanding of current problems and shortcomings

- Credibility gap

  Gap between the promise and reality of computing and data management support for microbial genome studies needs to be addressed, including

  Eliminate the disconnect between computer science R&D and bio research studies

  - Bio studies need “production” data management & computing infrastructure
  - Computer science R&D tends to produce illustrative “prototype” systems and applications that have little or no practical value for real-life biological studies